Follow Me to Unregulated Waters!

Daniel Holznagel 2024-05-30T11:37:12

The Digital Services Act (DSA) is aiming at making the internet safer. Amongst others, the DSA is empowering users to notify platforms about illegal content and (!) to make them take action—so called "notice and action", Art. 16 DSA. Just a few weeks ago, the European Commission opened proceedings against Meta concerning, amongst others, its reporting mechanisms. Obviously, the Commission is suspecting infringements by Meta in this field—though no details have been published as of today.

In this article, I will demonstrate how some major platforms are failing to properly implement the DSA's rules on notice and action mechanisms. In my view, many platforms are unduly nudging potential notice-senders (hereinafter: *reporters*) to submit weak, largely unregulated Community Standards flags. At the same time, platforms are deterring users from submitting (strong) notices regulated under the DSA.

For illustration, TikTok will serve as an example. However, many online platforms are showing similar flaws. E.g., Facebook's and Instagram's reporting mechanisms are as problematic as TikTok's.

The findings in this article are based on a collaboration with the Human Rights Organization HateAid, which has launched broad investigations into reporting mechanisms of all major platforms.

Parallel Reporting

For a better understanding I shall firstly explain the phenomenon of parallel reporting flows. Theoretically, platforms *could* take Art. 16 DSA as a floor, not as a ceiling, and treat *all* content moderation reporting according to the standards set out in Art. 16 DSA. However, all major platforms opted against this. Instead, they are channeling reporting into two then strictly separated sets of complaints:

- (mostly unregulated) complaints about alleged violations of Community Standards or Terms of Services (ToS), hereinafter: ToS-flag,
- (regulated) complaints about possible violations of EU- or Member States laws, hereinafter: DSA-notice.

To some extent, platforms can justify this separation with transparency obligations, as they are required to report on the metrics of DSA-notices. However, platforms might follow a purely economic rationale when, beyond differentiating for transparency reasons, they heavily aim at putting reports into two very distinct categories and treating them differently: DSA-notices must be processed under the regulatory umbrella and oversight of the DSA, while ToS-flags are much less regulated. On average, processing of DSA-notices requires more resources and might lead to stricter accountability. E.g., refusing to act on notified content after receiving DSA-notices might result in costly follow-on-measures through Art. 20, 21 DSA (which require human intervention). Moreover, only DSA-notices might trigger strong regulatory oversight, e.g. regarding Art. 16(6) S. 1 DSA which requires platforms to “process … notices … in a timely, diligent, non-arbitrary and objective manner”. Therefore, one can assume that platforms prefer to have reporters send them ToS-flags instead of DSA-notices.

And indeed, platforms are (unduly) designing their reporting mechanisms to foster such outcomes:

Illustration: TikTok’s Reporting Mechanism

Nowadays, most reporting of content takes place through pre-designed click-through reporting flows. Interestingly, major platforms opted for a pretty similar design of these reporting flows. Platform design here mainly consists of three relevant stages, which I refer to as “Initiation”, “Segregation” and “Submission”. Let me illustrate this for TikTok:

Step 1—<u>Initiation</u>

Users can enter the reporting process through symbols directly attributable to the piece of content in question (hereinafter “Initiation”-stage of reporting mechanisms). Platforms might use a flag-symbol, three dots, or—as illustrated here mimicking TikTok’s in-App-design—an arrow (which leads reporters to a submenu containing a flag-symbol):

99
username

(illustration: Initiation-stage with Arrow-link)

Step 2—<u>Segregation</u>

Clicking on the arrow and then (in the submenu) on “Report” will lead to a second, most consequential stage where reporters are asked to specify their concern (hereinafter: Segregation-stage), as shown in the next illustration simulating TikTok’s design:

(illustration: Segregation-stage)

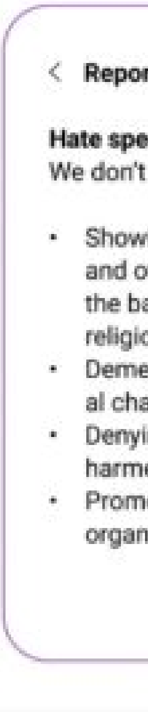
Reporters are asked to “select a reason” from a list of categories of prohibited content, for most parts mirroring the platform’s Community Standards. But that list also includes a link referring to “illegal content”, e.g. for TikTok: “Report illegal content”. For TikTok, note that, with some smartphones, it is necessary to scroll a bit to find the link “Report illegal content”.

From a platform perspective, this Segregation-stage is crucial, as it will determine whether the platform will treat the respective report as a mere ToS-flag or a DSA-notice. We must assume that platforms only treat reports as DSA-notices when reporters click “Report illegal content” (in the example). Transparency reporting according to Art. 15 DSA supports this assumption (reported metrics for DSA-notices support the conclusion; some platforms even provide corresponding explanation, e.g. TikTok’s DSA Transparency Report September to December 2023, p. 4: “… we have introduced an additional reporting channel … to ‘Report Illegal Content,# which enables users to alert us to content they believe breaches the law”).

Step 3–<u>Submission</u>

Following this Separation-stage, reporters are then taken to the final stage of the process, where they are asked to submit their report. However, this Submission-stage is heavily segregated, meaning that it looks very different for ToS-flags compared to DSA-notices.

If reporters had been clicking on one of the many descriptions for Community Standards violations (e.g. “Hate and harassment”), their (ToS) Submission-stage will look pretty neat and easy:



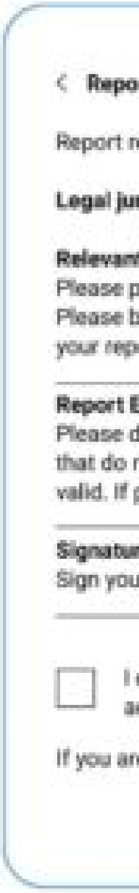
< Repor

Hate spe
We don’t

- Show and o the ba religio
- Deme al cha
- Denyi harm
- Prome organ

(illustration: Submission-stage for ToS-flags)

However, when, at prior Segregation-stage, reporters had been clicking “Report illegal content” leading platforms to treat the report as a DSA-notice, then things get complicated, and reporters are directed to a rather deterring Submission-stage, as (abstractly) illustrated for TikTok here:



< Repo

Report n

Legal ju

Relevant

Please p

Please b

your rep

Report E

Please d

that do n

valid. If p

Signatur

Sign you

□ I c
a

If you ar

(illustration: Submission-stage for DSA-notices)

In this case, reporters are obliged to specify “Legal jurisdiction”, and to provide an email address (this is only required for logged out users, not shown in the graphic), a “Report Explanation”, a “Signature” (“legal name”) and a confirmation of accuracy. Moreover, reporters are warned about possible consequences of unfounded reports (it requires scrolling to see the warning, not shown in the graphic above).

Reporting Mechanisms Not “User-Friendly”

In my view, the platform design illustrated above goes against the imperative of Art. 16(1) to put in place “easy to access and user-friendly” reporting mechanisms:

Initiation-Stage:

<u>Initiation not easy to access</u>: On TikTok’s app, the most conventional way to enter the reporting mechanism is through clicking the arrow-symbol, which will lead you

to a submenu containing a “Report”-link. However, the arrow’s main function is well-known for sharing content. Thus, users might not expect an entry to reporting. One might therefore argue that in TikTok’s app, reporting mechanisms are not “easy to access” (Art. 16(1) S. 2 DSA), as they are not “clearly identifiable” (as required by Recital 50 S. 3 DSA).

Segregation-Stage:

Misleading users towards uninformed decisions: At Segregation-stage, reporters might be misled about the consequences of the decision they are about to make. The platform is seemingly asking to specify the *nature of the reported content* (“Select a reason”). But without explicit explanation, the platform will use the decision to determine the *legal nature of the report* (weak ToS-flag or strong DSA-notice).

“Report illegal content” is difficult to find: It seems that only the link “Report illegal content” will make the platform treat a user’s reporting as a strong DSA-notice. However, it takes some effort to identify the relevant hyperlink (one link among many others, all indicating somewhat similar “reasons”) and scrolling might be required to find the link.

Nudging reporters against their legitimate interests: Both aforementioned factors implicate a substantial nudging towards (weak) ToS-flagging. This nudging comes at the cost of legitimate interests of users: Average reporters might not think about the differentiations between ToS-flags and DSA-notices at all. In doubt, platforms should assume that reporters will prefer to rely on a robust (and regulated) notification regime, that is, a (strong) Art. 16 DSA-notice. When platforms are trying to turn this logic upside-down by incentivizing ToS-flagging over DSA-notices, they are doing so only in their own economic interests and at the costs of average reporters’ legitimate interests.

Submission-Stage:

Illegitimate requests: While Art. 16(2) DSA requires platforms “to enable and to facilitate the submission of notices containing” granular information, Art. 16 does not (!) allow to make such data conditional (Recital 50 S. 5 DSA: “should allow, but not require”). Thus, TikTok is illegitimately requiring reporters to provide email address (logged out users), jurisdiction, explanation, and signature. This will deter users from submitting DSA-notices.

Burdensome and puzzling requests: TikTok’s request to specify a jurisdiction might also confuse reporters: What is TikTok referring to? A legal forum? Applicable speech restrictions? Most likely: the latter. However, in this case its menu lacks a check-box for “Union law”. Another shortcoming: Reporters cannot select multiple jurisdictions.

Disproportionate warnings: TikTok’s warnings about unfounded reports seem out of proportion. The warnings also seem unbalanced as TikTok does not provide respective warnings for ToS-flags.

Erroneous Processing of DSA-Notices?

Not only does such a platform design lead to a violation of Art. 16(1) DSA (reporting *mechanisms* not “easy to access and user-friendly”). In my view, it also leads to follow-on mistakes: As we have seen, platforms are nudging reporters to submit “weak” ToS-flags instead of “strong” DSA-notices. However, this only determines de-facto treatment of notices: From *the platforms’* perspectives, reports will mostly belong in the bucket of mere “unregulated” ToS-flags. But how platforms categorize a given report does not ultimately determine the true legal nature of that report.

Art. 16(1) S. 1 DSA legally defines what a DSA-notice is: A communication (“notify”) about a specific content that the reporter “considers to be illegal content”. This definition does not depend on how a platform *subjectively wants* to categorize a given report. Instead, the following question is decisive: How would a neutral third-party observer interpret the report? To answer this question, platform expectations must be considered, but only within the realm of “legitimate expectations” (a well-accepted principle for interpretation of legal declarations). That in mind, one can well argue that even when reports—through platform nudging and misleading—land in the ToS-bucket, the respective reporter (from a neutral observer’s perspective) might still aim at reporting content which *the reporter* considers “illegal”. His report then still is a DSA-notice, though the platforms might erroneously *think* they are allowed to treat it as mere ToS-flag, but this expectation is not legitimate. Think of the following example: A reporter does, at Segregation-stage, find and click “Report illegal content”. He then is scared by the unjustified requests for name, explanation, and jurisdiction. Discouraged, he clicks back to Segregation-stage and there clicks “Hate and harassment”, which TikTok then will handle as a mere ToS-flag. Would an observant third party necessarily interpret such reporting accordingly? I don’t think so!

If one follows this reasoning, then platforms will—in vast amounts—erroneously treat DSA-notices as mere ToS-flags. E.g., platforms will very likely not include these reports in their transparency reporting (Art. 15, 24 DSA), and they might not allow mandatory reviews of moderation decisions (Art. 20, 21 DSA). All these are follow-on mistakes infected by the platforms’ decision to one-sidedly steer reporters towards ToS-flags.

Conclusion and Outlook: DSA-Proceedings?

Through the design of their reporting flows, platforms are nudging users to submit weak ToS-notices, which leads platforms to count fewer (strong) DSA-notices falling under the regulatory oversight of the DSA. Such a design might be described as a “follow me to unregulated waters” – approach. In my view, this amounts to a violation of Art. 16(1) DSA. It also might lead to follow-on mistakes when DSA-notices are erroneously not treated as such.

In this article, TikTok just served as an example. All major platforms, to some degree, show similar shortcomings (Facebook, Instagram and X seem similar, LinkedIn seems worse, YouTube and Pinterest do better). As a promising start, the Commission has opened proceedings against Meta regarding its reporting mechanisms. And we might hope that the Commission will look into the aspects described here. But competent Digital Services Coordinators should also start investigating other platforms (all platforms mentioned have their seat in Ireland, putting Coimisiún na Meán in charge).
